DOI:
10.1016/j.future.2011.06.006

Document Version
Early version, also known as pre-print

Humanities e-Science: From systematic investigations to institutional infrastructures

Tobias Blanke[a], Mark Hedges[a]

[a]*King's College London, Centre for e-Research, 26-29 Drury Lane, London, WC2B 5RL, United Kingdom*

Abstract

In this article we bring together the results of a number of humanities e-research projects at King's College London. This programme of work was not carried out in an *ad hoc* manner, but was built on a rigorous methodological foundation, firstly by ensuring that the work was thoroughly grounded in the practice of humanities researchers (including 'digitally-aware' humanists), and secondly by analysing these practices in terms of 'scholarly primitives', basic activities common to research across humanities disciplines. The projects were then undertaken to provide systems and services that support various of these primitives, with a view to developing a research infrastructure constructed from these components, which may be regarded as a 'production line' for humanities research, supporting research activities from the creation of primary sources in digital form through to the publication of research outputs for discussion and re-use.

Keywords: humanities, eResearch, eScience, scholarly primitives, research infrastructures

1. Introduction

The programme outlined in this paper represents work being carried out by the Centre for e-Research[1] at King's College London. Part of the Centre's remit is to investigate and develop ICT infrastructure and tools for supporting and enhancing research practices across the institution, and, while this

Email addresses: tobias.blanke@kcl.ac.uk (Tobias Blanke), mark.hedges@kcl.ac.uk (Mark Hedges)

[1] www.kcl.ac.uk/iss/cerch/

remit is discipline-independent, there is a particular focus within the Centre on research in the humanities. This focus arises from the Centre's absorption of the former Arts and Humanities Data Service[2] and Arts and Humanities e-Science Support Centre[3], its ongoing collaborations with the Centre for Computing in the Humanities[4] at King's, and its participation in the EU ESFRI project DARIAH, which is developing a European research infrastructure for the humanities[5].

Building on experiences elsewhere in e-Science, our approach to infrastructure development was not based on a 'big bang', but rather on a bottom-up approach that involved the development of a number of smaller projects that addressed different aspects of the research lifecycle in the humanities (see also [1]). Some of these components have the characteristics of Virtual Research Environments (VREs) [2], by which we understand collaborative digital environments that facilitate the integration of information resources and tools to support a particular set of research activities. These activities ranged from very specific tasks, such as the creation of XML-based textual resources, through to much more general-purpose activities, such as the organisation and annotation of documents. Other projects focused on developing a service or tool to meet a single specific need. But in any case, the projects were developed with the ultimate goal of being able to provide a composite infrastructure to support the entire research lifecycle for the various humanities research communities across the institution, and by extension for their collaborators in other institutions.

The projects were not developed in an *ad hoc* manner, but were based on a rigorous methodological foundation. Firstly, we ensured that the work was thoroughly grounded in research practice by engaging with humanities researchers, looking at 'digitally-aware' activities as well as more traditional ones. Secondly, we analysed and classified these activities using a framework based on a set of 'scholarly primitives', that is to say basic activities that are common to research processes across humanities disciplines. The resulting model is a loosely-coupled composite of components, which may be regarded as a 'production line' in which the primary sources (in either physical or born-digital archives) that constitute the 'raw material' of research are processed

[2]http://www.ahds.ac.uk
[3]http://www.ahessc.ac.uk
[4]http://www.kcl.ac.uk/schools/humanities/depts/cch
[5]http://www.dariah.eu

through to research outputs that can be shared and discussed.

The paper is organised as follows: in Section 2, we outline the framework of 'scholarly primitives'. We use these to analyse and represent the research processes that we aim to support by means of our infrastructure; Sections 3 to 6 describe with reference to these primitives the projects that are providing the components from which the infrastructure is being constructed; and in Section 7, we show how the outputs of these projects can be linked together to form a broader environment for supporting the research lifecycle in the humanities.

2. Scholarly Primitives and Research Infrastructures

Traditionally, much humanities research was carried out on the basis of primary sources that were embodied physically in some form, either in a memory institution such as a museum, library or archive, or in the wider environment, such as buildings, archaeological remains or, indeed, people. For example, a scholar might visit some archives, search through them using whatever finding aids were available, find documents relevant to the topic in which they were interested, assemble them into various collections, and make notes on them. Something encountered in one document might lead them to another document, possibly in another archive, resulting in a chain of reading between resources, and ultimately, the scholar might produce a monograph or article.

Such traditional activities of humanities research translate naturally into the digital realm, and at King's College London we have been working for some time on how to best facilitate this translation while at the same time supplementing it with new methods that allow scholars to make use of the new opportunities that the digital realm offers. A number of projects have been developed to help us explore this translation. However, it soon became apparent that we needed to ensure that these projects did not remain isolated activities but rather came together as parts of an integrated whole, and to this end we found it useful to structure our work around the concept of 'scholarly primitives'.

Scholarly primitives may be defined as 'basic functions common to scholarly activity across disciplines' [3], and as such they can provide a conceptual framework for classifying scholarly activities and thus a firm foundation for conceptualising and developing an infrastructure in order to support these

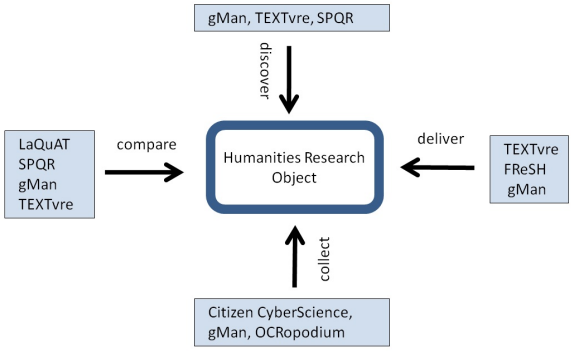

Figure 1: Primitives and Projects

activities. The concept has proven to be intuitive and valuable in multi-disciplinary endeavours such as humanities e-Science, and in particular for designing infrastructures that are sufficiently generic to cater for different research needs while not being engineered beyond the requirements and understanding of researchers.

An infrastructure based on such primitives will not have a purely disciplinary focus, but may be regarded as a marketplace or 'trading zone' [4] in which researchers can exchange and discuss their products and services. These products have changed significantly in recent years. While in the pre-digital world scholarly products were mainly articles and monographs, traded by means of the scholarly publishing industry, nowadays researchers produce a wide variety of digital outputs, not just (e-)publications but also databases, XML corpora or other online resources, which may be composed into more complex *research objects*, compound structures that bundle up related inter-related information objects of various forms [5]. In Figure 1, we recognise this shift of focus by placing the humanities research object at the centre.

Palmer et al. [3] define five scholarly primitives: searching, collecting, reading, writing and collaborating. It should be noted that this is one classification among several – see, for example, [6] or [7] for other analyses – but given the variety of humanities research perspectives this variety is to be expected, and is moreover not a problem, as the important thing is to provide a conceptual framework for anchoring our analysis. Indeed, while for the humanities in general a variety of activities, such as browsing, collecting, note taking, etc., have been identified as key components of everyday research, our

own objectives in identifying primitives were somewhat different – to provide a framework for developing research infrastructures – and so we reexamined these primitives in terms of their utility for achieving these objectives.

To ensure that our analysis was grounded in the needs of *researchers* rather than infrastructure developers, at all stages of this process we engaged with humanities researchers. We did this both within King's College London and, more widely, in the context of DARIAH, conducting semi-structured interviews in which questions were organised around viewpoints that examined how work is coordinated, planned and formalised. The focus groups of researchers included digitally-aware humanists, as we were interested not only in how traditional practices can be translated into digital ones, but also how these translations can be supplemented with techniques that have no traditional analogues.

We concluded that we needed a slightly different set of primitives. Some of Palmer's primitives do not really correspond to infrastructure work, for instance support for the activity of reading. At the same time, we added primitives particular to digital research in the humanities that might not appear as primitives in traditional humanities research. In [8], we summarised this work and identified five primitives relevant to infrastructure work. These are discovering, collecting, comparing, delivering and collaborating. Each of these in turn includes more fine-grained scholarly activities. For instance, discovering includes searching but also browsing and other more advanced techniques for 'finding out about' something.

In this paper we want to focus on the first four of these primitives, ignoring for the time being collaboration, as we would like to concentrate on demonstrating how humanities analysis can be improved for individual researchers. Figure 1 summarises how recent and current projects carried out at at King's College London support and influence scholarly activities that work with humanities research objects. In the following sections, we describes these projects in more detail in terms of the scholarly primitives to which they relate.

3. Discovering

As discovery plays such a central role in any humanities research activity, we could classify almost all of our projects under the *discovering* primitive, but instead we will concentrate on a project that brings a new perspec-

tive to the issues of discovery in humanities research: gMan[6], which offers researchers highly customisable discovery services, and indeed goes beyond that by pioneering a generic analysis environment for collaborative, data-driven research in the humanities [9].

Our user engagement activities demonstrated a need among humanities researchers for a general-purpose and scalable environment that supports the on-demand integration of *ad hoc* collections of heterogenous and dispersed data sources, and is provided with effective services for discovery across those resources. An additional challenge here is that the on-demand deployment of efficient discovery mechanisms requires an infrastructure that supports the on-demand indexing of potentially large numbers of document sets.

For implementing such an environment, our starting point was *D4Science*[7], a production-level infrastructure that mainly serves scientific communities, but which is not biased towards any particular discipline and has great potential for meeting these requirements for integration and discovery. *gCube*[8], on which the infrastructure is based, was originally developed in the context of the European infrastructure project DILIGENT [10], which aimed to create a grid-based digital library system. It is a distributed, service-based system designed to support the full life-cycle of modern research, with particular emphasis on application-level requirements for information and knowledge management [11], and to this end gCube interfaces with European grid middleware and research infrastructures to exploit shared access to computational and storage resources.

At the centre of gMan lies not just the data but also the functionality for working with that data – services that collate, import, describe, annotate, merge, transform, index, search and present information for various multidisciplinary communities. In respect of the *discovering* primitive, gCube supports several search and browse services, which are supported by a variety of indexes that can be created on demand by users. The researcher can perform a text-centric search across any collections that are available in his or her work area, where by *text-centric* search we understand a search across documents that copes with uncertainty and non-exact matching, and produces a ranked result set, analogous to how a search engine deals with

[6]http://gman.cerch.kcl.ac.uk
[7]http://www.d4science.eu
[8]http://www.gcube-system.org

the Web. In this way, way the researcher is helped to find relevant resources, filter and select from them, and use them as the basis for further searches. As well as full-text searches, discovery can also be based on the metadata records associated with datasets, on XML elements within XML-encoded datasets, or on geo-spatial attributes. gMan is able to identify common XML fields across heterogeneous data sets, and users can search across these common fields only, thus comparing results across the data sets. This activity of *comparing* now leads us to consider our second scholarly primitive.

4. Comparing

Comparing digital resources generally requires the existence of some degree of commonality across the resources, a frame of reference within which they can be compared, as for example when they are guided by similar standards. However, while the development of such standards is of course important, it will not solve all issues raised by comparing data in the humanities. For one thing, there exists a great deal of legacy data in diverse formats. Moreover, even when standards are used, the sheer variety of humanities data and research means that there is a great deal of variation in how these standards are applied. More importantly, however, standards are generally developed within particular domains, whereas research is often inter-disciplinary, making use of varied materials, and incorporating data conforming to different standards. There will inevitably be diversity of representation when information is gathered together from different domains and for different purposes, and consequently there will always be a need to integrate diverse representations so that information can be compared across them.

In this section, we describe a number of projects we have undertaken in recent years that address activities of *comparing* across heterogenous resources. We begin with our attempt in the LaQuAT project to use grid-based technologies for integrating humanities archives. We discuss the shortcomings of this approach and show how we addressed them in the subsequent SPQR project by using Linked Data for providing integrated access. But comparing is not just about integrating; it also includes activities such as annotating, as annotations can be regarded as one of the outcomes of *comparing* activities.

The LaQuAT (Linking and Querying of Ancient Texts) project[9], funded by JISC via the ENGAGE programme in 2007, tried to overcome the lack of

[9]http://laquat.cerch.kcl.ac.uk

standards across humanities resources by developing data integration workflows based on the OGSA-DAI software [12]. OGSA-DAI can be used to make heterogeneous datasets appear as a single virtual data resource that can be queried via a standard SQL-like interface, enabling the user to compare even highly heterogenous resources. However, while this integration was achievable technologically, when it was evaluated by humanities researchers they raised significant issues. Comparing datasets in this manner required a significant understanding of the underlying semantics of the data at a fine-grained level, semantics that were for the most part left implicit in these relational databases, and were complicated further by the variety of conventions used in representing data. For example, dates may be given in very different forms, and may be expressed with very different precisions and levels of confidence. The fuzzy, uncertain and interpretative nature of the available data also confused the semantics of integration, and made it difficult to describe the relationships between data sets. It was not always clear whether similarly named columns in independent databases really represented the same sort of information and could validly be linked. In some cases there were deeper semantic issues, for example when two independent datasets contained contradictory information.

Another major issues with the LaQuAT approach was that datasets remain as isolated silos, albeit accessible from a single place. Additional relationships identified by researchers as a result of *comparing* activities cannot be explored. We considered that Semantic Web and Linked Data approaches have great potential here, as they allow researchers to formalise resources and the links between them more flexibly, and to create, explore and query these linked resources. Closely allied to Linked Data has been work on ontologies for providing agreed meanings for both links and the resources they connect. Ontologies can thus act as the semantic mediator between heterogeneous datasets, enabling researchers to explore, understand and extend these datasets more productively and so improve the contributions that the data can make to their research. To investigate the possibilities of Linked Data for comparing humanities datasets, we have recently started a new JISC-funded project called SPQR[10], in which we will investigate the use of the recently released Europeana Data Model [13] as an integration ontology and mediator for heterogeneous humanities datasets. This approach has the added advan-

[10]http://spqr.cerch.kcl.ac.uk

tage of facilitating the publication of these datasets through the Europeana portal.

SPQR investigates exposing humanities datasets using RDF as a basis for exploring and comparing them. The ultimate objective will be to bring the transformed information into a common corpus or 'RDF warehouse', where it can be explored and searched in an integrated way, and where new comparisons and connections (corresponding to new RDF or similar statements) can be made by the researcher and added to the corpus of information. We will investigate mechanisms for breaking this information out of its current silos, and transforming it from its legacy formats (such as databases) into our chosen representation, and exposing it as Linked Data. There are two broad approaches to transforming datasets in this way – using wrappers for on-the-fly conversion, and converting data before exposing it – and we evaluate the pros and cons of each.

As the work progressed it became clear that the activities of *integrating* resources and *annotating* resources are intrinsically linked, as most resources in the humanities cannot be integrated automatically but require human input for establishing links between them. Indeed, in Palmer's classification [3], *annotating* activities form a subset of those falling under the *comparing* primitive. We have just seen how annotations can be used to enhance existing information resources by creating new links, and thus in turn to support further scholarly *comparing* activities. To clarify further the relationship between comparing and annotating, we now revisit gMan and TEXTvre as examples of, respectively, general-purpose and highly specialised annotation environments.

The gCube environment provides services to support comparing and annotating of datasets. A textual note can be attached to an object (or to part of an object) in gCube, and similarly a labelled link can be created created between objects, in both cases the annotation being marked with the timestamp and the user who created it. Thus gMan supports general-purpose annotating activities.

In contrast, digital research in the humanities may require highly specialised annotation services; for example, the central scholarly activity supported by the TEXTvre environment may be regarded in this way, as TEXTvre concentrates on supporting the creation of deep annotations in the TEI stan-

dard for encoding digital texts[11]. TEXTvre aims to create a unified environment, embedded within institutional research practices at King's College London, that supports digital humanities researchers in the creation of online critical editions based on TEI/XML resources. The project builds upon the success of the German TextGrid[12] project and reuses much of its technologies and methodologies.

While a manual approach to annotation of texts is an essential requirement for humanities researchers, it is labour-intensive and not scalable as the quantity of digital material increases. On the other hand, as larger corpora of texts are built up, large-scale annotations make possible qualitatively different forms of scholarly activity. Consequently, the architectures of TEXTvre and its parent TextGrid support the (partial) automation of annotation work, in the case of TEXTvre by integrating the information extraction and text mining facilities provided by the GATE environment[13]. GATE is a language engineering toolkit with services for the automated processing, analysis and visualisation of documents, and it provides an environment in which pipelines of Natural Language Processing tools, such as tokenisers, part-of-speech taggers and parsers, can be run over a corpus of documents to create sets of annotations. In TEXTvre, GATE supports the annotation process by providing the researcher with various standardised suggestions for annotation items.

5. Collecting

Almost all our systems support *collecting* digital content in some fashion. gMan, for example, supports the collecting together of objects from dispersed and heterogenous digital resources into collections, either by means of static membership list or dynamically by specifying membership criteria, which can then be manipulated as objects in themselves – a researcher can refer to it by an identifier, search across it, and share it with colleagues. We term such objects *virtual collections*, and, while conceptually straightforward, the ability to build such collections is very important for allowing researchers to deal with large quantities of primary documents and other data.

[11] http://www.tei-c.org/index.xml
[12] http://textgrid.de
[13] http://gate.ac.uk

This still depends, however, on using *existing* metadata provided either by the original creators of the digital content or by other parties that have enhanced that content since its creation. In this section, we concentrate rather on collecting in the sense of building collections, whether these are collections for long-term use, such as the critical scholarly editions with which TEXTvre is dealing, or more short-term collections to support a particular research task or for writing a paper.

Collecting complex digital content in repositories and related systems is a field that is both well-researched and widely practised (see, e.g., [14]). However, the creation of the contextual information or metadata required to discover and use this content, or even to understand it, raises significant practical issues of scalability, as in general this work requires professional staff with particular domain knowledge. Here we consider two projects that investigate alternative mechanisms for collecting information about digital resources, one using communities of volunteers for building descriptive metadata, the other using open source optical character recognition.

In the community approach, targeted volunteers associate themselves with an archive in order to maintain and enhance it for subsequent use. In principle, there are already widely used models for a community-based approach to enhancing digital resources on the web. Currently, the predominant way of building community-based metadata is to use free tagging, where the public is encouraged to describe or 'tag' objects, in a way analogous to the social tagging used on sites such as Flickr and similar websites. While this may be useful in the context of these social websites, it is be too libertarian and open to misuse to produce results that are acceptable in more formal cultural heritage environments. For example, the 'steve' project[14], a collaboration between cultural heritage organisations in the USA, tried this approach with rather mixed results, finding that the terminology used for tagging was too loosely connected with the sort of terms used by the cultural heritage community to form an effective basis for searching.

Instead of this we have been investigating a 'volunteer thinking' or 'citizen cyberscience' approach to enhancing archives. The idea behind volunteer thinking is that a body of work is split up into small, self-contained tasks, which can be assigned to volunteers and performed over the internet. We

[14]http://www.steve.museum/

use the open-source software Bossa[15], which allows groups of people to participate online in the organisation and cataloguing of collections. Bossa is designed to harness the 'distributed thinking' of participants by allowing us to define policies and tasks that guide the users in their decisions. Distributed thinking has proved highly reliable to support the collecting of information about resources because the process incorporates highly redundant verification of results from many volunteers. More subtly, it also adapts to variations in ability, maintaining an estimate of the skill of each participant, and ensuring that there is a sufficient consensus of redundant results among an appropriate set of volunteers.

We have been demonstrating this with the East London Theatre Archive (ELTA)[16] project, which had previously digitised a variety of material such as playbills and programmes relating to theatres in the East End of London in England, material which originated from a number of physical archives. In many cases these archives were relatively small and inaccessible to researchers. While the project made a significant number of images and metadata available online, the digitised images contain a great deal of information that has not been extracted and cannot be utilised in searches.

Volunteers can rectify this situation, and the nature of the material enables us to involve quite different sorts of community: on the one hand, theatrical communities, and on the other hand, local communities, including schools and local and family history groups. Tasks performed to enrich the resource include: dividing up complex images into distinct sections, such as advertisements, lists of performers, and so forth; transcriptions of text from the images; linking textual components with thesauri, such as performers, thus producing a resource that is connected with the wider online world of data.

We are integrating this approach with other components of the institutional environment in two ways. Firstly, with the institutional repository infrastructure, which is used to curate and preserve a variety of digital material; enhancements acquired though community participation are also managed by the repository, together with semantic and other relationships between the objects, and (crucially) provenance metadata for the information captured. It is also being integrated with the results of another project,

[15]http://bossa.berkeley.edu/
[16]http://www.elta-project.org/

OCRopodium [15], which is creating OCR workflows for historical collections, using the open source OCRopus software. Although a major part of this project concerns training OCR software to produce improved results on difficult or complex (from an OCR point of view) texts, quality can be improved by integrating the services developed into workflows that include human interventions. The ELTA project is making use of this to use community participation for improving and correcting initial OCR outputs.

A key infrastructure component for supporting *collecting*, at least for more persistent collections (in contrast to virtual collections created temporarily), is the institutional repository infrastructure. This is based on the Fedora Commons software[17], which specifically aims at managing complex digital resources with interrelationships between objects [14]. While we have been investigating and applying Fedora for some time, and not only in the humanities, we are currently undertaking a project CMES[18], funded by the Arts and Humanities Research Council[19], which is taking a systematic approach to the development of *patterns* for digital humanities content. Representations of digital objects within Fedora are formalised as *content models*, and our content patterns will be based upon these, and in particular upon the approach developed by the State and University Library of Denmark[20], enabling us to develop a consistent, flexible and modular framework for representing digital humanities content. Use of the repository as a common linking factor brings other advantages, facilitating on the one hand interoperability between the various components of the emerging infrastructure, and on the other integration with institutional digital preservation policies and systems [16].

6. Delivering

In this section, we consider how various modes of *delivering* of humanities resources are supported by our emerging infrastructure. Three modes are considered: the Web-based publication of stable digital humanities resources; the interim sharing of temporary research results within a particular research community; and the exposure of humanities research objects in machine-readable form for use by software agents.

[17] http://fedora-commons.org/
[18] http://cmes.cerch.kcl.ac.uk
[19] http://www.ahrc.ac.uk
[20] http://sourceforge.net/apps/mediawiki/ecm

Integration with the College's repository infrastructure is key to the success of the TEXTvre project, which was driven by the use case of creating and publishing scholarly critical editions. On the one hand, the lifecycle of the diverse digital material used and produced by textual scholars can result in complex semantic networks of digital objects, and integrating the TEXTvre/TextGrid platform with the repository helps to manage that complexity, as well as removing the need for implementing separate delivery mechanisms for the individual resources. As described in Section 5, digital collections are represented in this repository using a range of *content patterns*, and corresponding to these there is a framework of Web delivery components that are driven by the underlying content patterns. This has the benefit that these components are available for any resource that implements the corresponding patterns, leading to more consistent, interoperable and sustainable delivery mechanisms for these resources.

TEXTvre also allows for the controlled publication of work-in-progress into restricted repository spaces that are exclusive to individuals or small groups, before the resource is actually made public. As we saw in Section 5, gMan offers similar services, supporting the creation of *virtual collections*, which provide a mechanism allowing a researcher to share their work, including the relevant research material, annotations and links, with selected colleagues, who in turn add their own annotations and links that may confirm, extend or contest the researcher's conclusions. A researcher has full control over whether to keep results private, deliver them to a broader group, or indeed make them entirely public. In this way a scholarly dialogue is created and recorded. This could also facilitate new forms of publishing in the humanities, in which readers would have access to the reasoning process that lies behind conclusions, enabling them to validate it – the acceptance of humanities research often depends on provenance of information and on peer assessment – and perhaps criticise it.

Finally, let us consider the delivery of humanities research objects for consumption by software rather than human agents. The FReSH project [17] showed how combinations of standard web technologies such as ReSTful services and ATOM feeds can be used to deliver effective text mining services to end users in the Humanities. In earlier work, we had focused on the delivery of such services to the human creators of humanities websites so that the results of text mining could be accessed as easily as possible. However, software agents in the worldwide digital ecosystem are not well served by these human-centric representations, and so we concentrated on how to make

the information delivered by the FReSH text mining agent more machine-readable. To this end, we developed new digital surrogates for publishing to three categories of software agent on the Web: Google, text mining agents and the emerging agents consuming Linked Data.

A good example of the overall idea is the publication of indexing results for text mining agents. Typically, a text mining system is based on two fundamental processing steps [18]. In the first so-called indexing step, it aims to find a representation that models the available information content of the documents under consideration as optimally as possible, while at the same time delivering a representation that is computationally viable. In the second step, when the actual mining takes place, the system analyses the sets of documents presented to it in order to extract and discover information. These might be new relationships, such as previously unknown links between documents, or new facts. The indexing step is often based on extracting term frequency (TF) and inverted document frequency (IDF) to evaluate how much information a word contains for the purposes of identifying facts in a document collection. The assumption, that a word is more important the more it occurs in a collection, offset by the number of documents in which it occurs, has proven to be a powerful one. This TF-IDF information is collected in indexes and used for the actual analysis step. In FReSH, we decided that it often suffices to deliver the TF-IDF index representation of a document for other text mining agents to carry out their analysis, and we verified this assumption by publishing these to a test system that was able to create links to collections of secondary literature in JSTOR[21].

In discussing the LaQuAT project in Section 4, we noted that delivering structured data in formats such as relational databases, spreadsheets or XML files can make it very difficult for researchers to explore the datasets in an integrated fashion and to understand and exploit the connections between them. It is in precisely this regard that semantic approaches have great potential, as they allow researchers to formalise resources and the links between them more flexibly, and to create, explore and query these linked resources. Consequently, we were led to explore mechanisms for breaking the information out of such datasets and delivering it as Linked Data, and a number of tools have been explored for extracting information from other formats and delivering it as RDF, as well as for extracting entities such as

[21] http://dfr.jstor.org/

place and personal names from unstructured text and for proposing potential links between entities. The aim of the SPQR project has been to investigate the potential of this for the humanities, using datasets relating to ancient history and archaeology as test cases.

7. Producing Humanities Research Objects

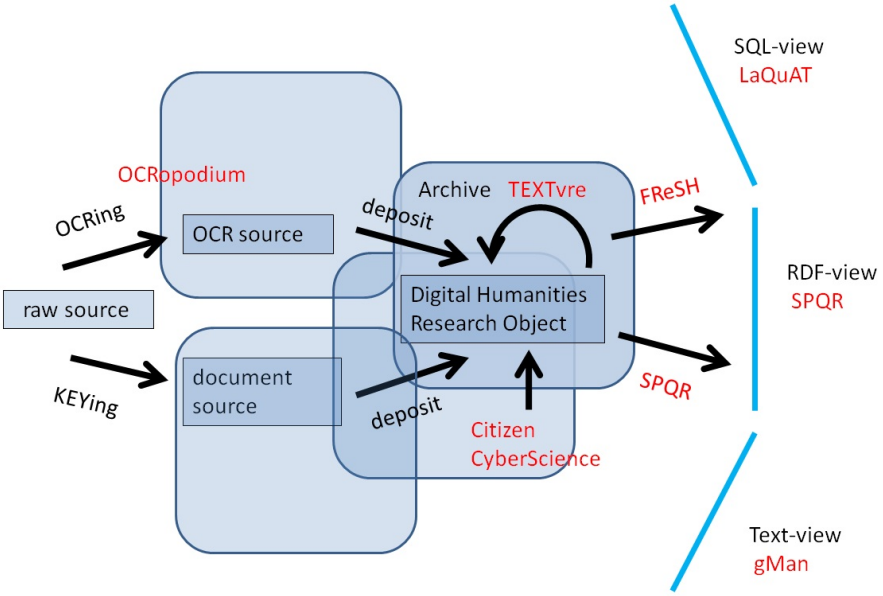

Figure 2: Producing humanities research objects

We now consider how these projects can be combined as distinct but integrated building blocks for the production of humanities research objects.

Our user engagement activities (see Section 2) gave rise to a view of research activities in much of the humanities as being complex and highly interactive workflows with the researcher at the centre. Researchers seek resources relevant to their interests; they select, interpret and analyse them, using tools but also their own judgement based on other available evidence both internal and external to the resource. The results of one search may, taken together with other information available to a researcher, influence the questions that are asked of others. Importantly, this may be a collaborative

activity, requiring the ability to record interpretation and opinion as annotations, and thus start a dialogue within the relevant community. However, the use cases can be very varied and unpredictable. They are 'workflows' in the sense that they comprise a sequence (or directed graph) of activities, in which the outputs of some activities form the input to others, but except in relatively specialised cases we rarely encountered workflows that could be automated, shared with and used by others, such as occur in many scientific disciplines [5]. Typically, the key controlling agent in a humanities workflow is the judgement of the individual researcher.

In much of the humanities, these research processes and dialogues have frequently had their starting point in archives of some form, that is to say in dedicated, and often specialist, collections of documentary material in various media that form the researcher's primary sources, the raw materials of scholarship [19]. In recent years there have been significant digitisation programmes for archival material, which to an increasing extent have been able to provide the researchers with easily accessible, digital surrogates for the physical archives, creating new possibilities for the scholar. Memory institutions and institutional archivists have been systematically digitising the source material for which they are responsible, and assembling it into broad digital collections representative of their holdings. On a smaller scale, individual humanists or small groups may need to digitise a more restricted body of material, quite possibly assembled from a number of different physical sources, that allows them to address specific research questions. While this form of *collecting* may seem to be qualitatively different, it is nevertheless an integral part of humanities research, requiring scholarly input in the selection and organisation of material, and the creation of metadata and other contextual information.

As we have seen, our approach to developing support for these research communities has been to decompose the use cases into common activities, using a conceptual framework based on scholarly primitives. These activities and primitives correspond (although not in a simple one-to-one fashion) to projects that we have undertaken to implement systems or services that support these activities, and which we are connecting up to produce a composite infrastructure or 'production line' for humanities research. 'Real' research scenarios can be modelled in terms of combinations of these primitives, which in implementation terms correspond to different pathways of usage through the infrastructure. This correspondence is illustrated in two ways: in schematic form in Figure 2; in Table 1, by outlining steps from a

concrete scenario of a scholar who is working with ancient documents, such as papyri or inscriptions, and showing how these steps map firstly onto our classification into scholarly primitives and secondly onto infrastructure components (see [9] for additional background to this example).

8. Conclusions and Future Work

In this paper we have outlined a programme for developing an institutionally-based infrastructure for supporting humanities research by identifying and implementing smaller projects that address different parts of the humanities research lifecycle, and which can be connected together to create a broader environment. These projects were identified not by chance, but by a rigorous approach to describing research practices that involved close engagement with humanities (and digital humanities) researchers, the analysis of these practices into core activities, and the classification of these activities using a framework of 'scholarly primitives'. The resulting infrastructure model is a loosely-coupled 'production line', in which archivists (and others) can transform (where necessary) the raw materials of humanities research into digital collections, and scholars can work and produce research objects that can be shared, discussed and re-used.

We have concluded from our work so far that the 'scholarly primitives' model is an effective way of giving structure to the diverse, unpredictable and user-centric workflows that form the day-to-day activities of many humanities researchers, particularly those that work with archives of one form or another. This model of what researchers do maps quite naturally onto a model of an infrastructure that supports these research processes. Naturally, the programme we describe continues to evolve as new projects are developed. To date, we have started to join up these projects by implementing and evaluating a number of short scenarios that combine multiple primitives; our next step is to carry out a number of more realistic and increasingly complex scenarios that are based on our interviews with researchers and which are representative of certain aspects of research scholarship.

Many of the activities addressed by these projects are shared with other disciplines, which raises two complementary questions for further investigation. To what extent can we transfer humanities models and modes of interpretation outside the humanities, for example in science and medicine, and what elements are intrinsic to the humanities? To answer these, we need to carry out a deeper investigation into the relationship between research

Scenario Activity	Primitives	Infrastructure
Create digital surrogates (typically XML-based) of physical documents.	collecting	TEXTvre
Create explicit collections of digital objects in formally managed and sustainable digital repositories.	collecting	CMES, volunteer thinking, OCRopodium
Look for related resources of various forms (documents, data, articles) created by other researchers.	discovering	TEXTvre, gMan, CMES, external web searches
Collect together resources selected from those discovered, forming a *virtual collection*.	collecting	gMan
Perform text-centric queries across a virtual collection.	discovery	gMan
Perform complex queries (e.g. date-based, geo-spatial) across a virtual collection.	discovery	gMan, SPQR, TEXTvre
Annotate an object (e.g. XML file, image, database row) or part of an object, for example to identify a person mentioned in a text, or to correct or dispute existing information.	comparing	gMan, SPQR, TEXTvre
Add a link between two objects, for example saying that one papyrus is in the same handwriting as another.	comparing	gMan, SPQR
Generate suggested links/annotations automatically.	comparing	SPQR, TEXTvre
Create research objects that package together relate research material, annotations and links	collecting	gMan, CMES, TEXTvre
Share research objects with selected colleagues, who in turn add their own input.	delivery, collaborating	gMan, TEXTvre
Publishing research objects and conclusions more widely.	delivery	CMES, SPQR

Table 1: Correspondence between scenario, primitives and infrastructure components

models and e-research technologies, building on the above analysis in terms of scholarly primitives.

The broader aim of our work is to develop and evaluate a means of providing 'whole lifecycle' research environments for communities in a variety of humanities domains, particularly those involved in archival work. Our work to date, which remains ongoing, will provide a solid basis for this by allowing us to roll out in phases an environment that exploits existing, lower-level infrastructures such as national grid infrastructures, and which can be used and evaluated by researchers within our institution and its collaborators, and, via the DARIAH community, across Europe. A key part of this will be an evaluation of the uptake of the infrastructures and resources made available, and of their impact on the research carried out.

[1] T. Blanke, M. Hedges, S. Dunn, Arts and humanities e-science: Current practices and future challenges, Future Generation Computer Systems 25 (4) (2009) 474–480. doi:10.1016/j.future.2008.10.004.

[2] M. Fraser, Virtual research environments: Overview and activity, Ariadne (44).

[3] C. L. Palmer, L. C. Teffeau, C. M. Pirmann, Scholarly information practices in the online environment: Themes from the literature and implications for library service development.
URL http://www.oclc.org/research/publications/library/2009/2009-02.pdf

[4] P. Galison, Image and logic: A material culture of microphysics, University of Chicago Press, Chicago, 1997.

[5] D. De Roure, C. Goble, R. Stevens, The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows, Future Generation Computer Systems 25 (2009) 561–567. doi:10.1016/j.future.2008.06.010.

[6] J. Unsworth, Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this?, Symposium on "Humanities Computing: formal methods, experimental practice" at King's College London (2000).
URL http://www3.isrl.illinois.edu/ unsworth/Kings.5-00/primitives.html

[7] A. Benardou, P. Constantopoulos, C. Dallas, D. Gavrilis, Understanding the information requirements of arts and humanities scholarship, International Journal of Digital Curation 5 (1).
URL http://ijdc.net/index.php/ijdc/article/view/144/0

[8] S. Anderson, T. Blanke, S. Dunn, Methodological commons: arts and humanities e-science fundamentals, Phil. Trans. R. Soc. A 368 (1925).

[9] T. Blanke, L. Candela, M. Hedges, M. Priddy, F. Simeoni, Deploying General-Purpose Virtual Research Environments for Humanities Research, Philosophical Transactions of the Royal Society A 368 (1925) (2010) 3813–3828. doi:10.1098/rsta.2010.0167.

[10] D. Castelli, L. Candela, P. Pagano, M. Simi, Diligent: a digital library infrastructure for supporting joint research, in: Local to Global Data Interoperability - Challenges and Technologies, 2005, IEEE Computer Society, Washington, DC, 2005, pp. 56–59. doi:10.1109/LGDI.2005.1612465.

[11] L. Candela, D. Castelli, P. Pagano, gCube: A Service-Oriented Application Framework on the Grid, ERCIM News (72) (2008) 48–49.
URL http://ercim-news.ercim.org/content/view/315/500/

[12] M. Jackson, M. Antonioletti, T. Blanke, G. Bodard, M. Hedges, A. Hume, S. Rajbhandari, Building bridges between islands of data — an investigation into distributed data management in the humanities, in: Proceedings of the Fifth IEEE International Conference on e-Science, IEEE Computer Society, Washington, DC, USA, 2009.

[13] C. Meghini, M. Doerr, A. Isaac, Definition of the Europeana Data Model elements, v. 4.11 (Feb. 2010).

[14] C. Lagoze, S. Payette, E. Shin, C. Wilper, Fedora: An Architecture for Complex Objects and Their Relationships, International Journal on Digital Libraries 6 (2).
URL http://arxiv.org/abs/cs.DL/0501012

[15] M. Bryant, T. Blanke, M. Hedges, R. Palmer, Open source historical ocr in the ocropodium project, in: Proceedings of European Conference Research and Advanced Technology for Digital Libraries (ECDL), 2010.

[16] M. Hedges, T. Blanke, A. Hasan, Rule-based curation and preservation of data, Future Generation Computer Systems 25 (4) (2009) 446–452. doi:10.1016/j.future.2008.10.003.

[17] T. Blanke, M. Hedges, R. Palmer, Restful services for e-humanities, in: 3rd IEEE International Conference on Digital Ecosystems, IEEE Computer Society, Washington, DC, USA, 2009.

[18] M. W. Berry, M. Castellanos, Survey of Text Mining II: Clustering, Classification, and Retrieval, 2007.

[19] W. Duff, B. Craig, J. Cherry, Historians use of archival sources: Promises and pitfalls of the digital age, The Public Historian 26 (2) (2004) 7–22. doi:10.1525/tph.2004.26.2.7.